MGSC 695 - Optimization for Data Science

Winter 2024

Homework Assignment 1

Submitted by: Jared Balakrishnan

McGill ID: 261175926

Problem 1

It is said that we are given data that consists of n observations. Understandably, there will be n predictions, which in turn implies that the matrix representing the outcome (or "target") variable will have n observations.

This outcome variable matrix is a $(n \times 1)$ matrix written as:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Similarly, the matrix representing the observations can be represented by a $n \times (k+1)$ matrix. x_{ij} represents the i^{th} observation of the j^{th} predictor.

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

And the k coefficients for each of the predictors, including one for the intercept can be written as a column vector with $(k+1) \times 1$ rows as follows:

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

With ϵ representing a $n \times 1$ column vector of errors (one associated with each of the n predictions), the multiple linear regression model can be written in matrix form as:

$$Y = X\beta + \epsilon$$

Sub-Problem 1

According to the rules of matrix multiplication:

- The number of columns in the first matrix should be equal to the number of rows in the second matrix.
- The resulting matrix will have the number of rows of the first matrix and the number of columns of the second matrix.

Therefore, from the relationship

$$Y = X\beta + \epsilon$$

Since X is a $n \times (k+1)$ matrix and β is a $(k+1) \times 1$ column vector, the term $X\beta$ would result in a $n \times 1$ column vector. Additionally, ϵ is a $n \times 1$ column vector.

According to the rules of matrix addition, two matrices can be added only if they possess the same order (same number of rows and columns).

Both $X\beta$ and ϵ are $n \times 1$ column vectors, as is the outcome variable matrix Y, thereby confirming that the dimensions of the matrices on both sides of equation 1 match.

Sub-Problem 2

The residuals are given by $e = Y - X\beta$

This implies:

$$e = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

which can be further simplified as:

$$e = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \beta_0 + x_{11}\beta_1 + \cdots + x_{1k}\beta_k \\ \beta_0 + x_{21}\beta_1 + \cdots + x_{2k}\beta_k \\ \vdots \\ \beta_0 + x_{n1}\beta_1 + \cdots + x_{nk}\beta_k \end{pmatrix}$$

$$e = \begin{pmatrix} y_1 - \beta_0 - x_{11}\beta_1 - \cdots - x_{1k}\beta_k \\ y_2 - \beta_0 - x_{21}\beta_1 - \cdots - x_{2k}\beta_k \\ \vdots \\ y_n - \beta_0 - x_{n1}\beta_1 - \cdots - x_{nk}\beta_k \end{pmatrix}$$

The sum of the squared residuals is given by $\mathrm{e}^{\mathrm{T}}\mathrm{e}$:

e^{T}, the transpose of e, a $1 \times n$ row vector is given by:

$$e^T = \begin{pmatrix} y_1 - \beta_0 - x_{11}\beta_1 - \cdots - x_{1k}\beta_k & y_2 - \beta_0 - x_{21}\beta_1 - \cdots - x_{2k}\beta_k & \cdots & y_n - \beta_0 - x_{n1}\beta_1 - \cdots - x_{nk}\beta_k \end{pmatrix}$$

Subsequently, the dot product operation $\mathrm{e}^{\mathrm{T}}\mathrm{e}$ yields a scalar s given by:

$$s = e^T \cdot e = (y_1 - \beta_0 - x_{11}\beta_1 - \cdots - x_{1k}\beta_k)^2 + (y_2 - \beta_0 - x_{21}\beta_1 - \cdots + x_{2k}\beta_k)^2 - \cdots + (y_n - \beta_0 - x_{n1}\beta_1 - \cdots - x_{nk}\beta_k)^2$$

We now estimate the parameters for the vector β (represented by β^*)by computing the first derivative of the scalar s with respect to each of $\beta_1, \beta_2, \cdots, \beta_k$:

$$\beta^* = \nabla s = \begin{pmatrix} \frac{\partial s}{\partial \beta_0} \\ \frac{\partial s}{\partial \beta_1} \\ \vdots \\ \frac{\partial s}{\partial \beta_k} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ik}\beta_k) \\ -2\sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ik}\beta_k) \\ \vdots \\ -2\sum_{i=1}^{n} x_{ik}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ik}\beta_k) \end{pmatrix}$$

Sub-Problem 3

Upon setting the derivative of the scalar **s** from sub-problem 2 to 0, we get the following:

For β_0:

$$-2\sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ik}\beta_k) = 0$$

This now becomes:

$$\sum_{i=1}^{n} y_i = n\beta_0 + \sum_{i=1}^{n} x_{i1}\beta_1 + \cdots + \sum_{i=1}^{n} x_{ik}\beta_k$$

For β_j where (j = 1, 2, ..., k):

$$-2\sum_{i=1}^{n} x_{ij}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ik}\beta_k) = 0$$

This now becomes:

$$\sum_{i=1}^{n} x_{ij}y_i = \sum_{i=1}^{n} x_{ij}\beta_0 + \sum_{i=1}^{n} x_{ij}x_{i1}\beta_1 + \cdots + \sum_{i=1}^{n} x_{ij}x_{ik}\beta_k$$

The LHS represents the product of the transpose of matrix X and column vector Y, giving it the form of $X^T y$. This results in a $(k + 1) \times 1$ vector.

The RHS represents the product of matrices X^T and X,before being multiplied by the vector β thereby giving it the form of $X^T X\beta$. This results in a $(k+1)\times 1$ column vector where each element is a linear combination of the beta values.

In matrix notation therefore, the LHS and RHS can be condensed into:

$$X^T y = X^T X\beta$$

If you found that explanation a bit too unwieldy, below shown is an alternate proof that uses matrix algebra and matrix calculus.

We are given that the residuals are given by $e = Y - X\beta$, and that the sum of squared residuals is given by $e^T e$.

The $(k+1)$ parameters in β are to be estimated by minimizing the sum of squared residuals (represented below by s), which involves taking its first derivative and setting it to zero.

Therefore,

$$s = e^T e$$

$$s = (Y - X\beta)^T (Y - X\beta)$$

By the rules of matrix algebra, we can now write:

$$s = (Y^T - X^T\beta^T)(Y - X\beta)$$

Multiplication is distributive:

$$s = Y^TY - Y^TX\beta - X^T\beta^TY + X^T\beta^TX\beta$$

From the above expression, we can see that the second and third terms are transposes of each other. That is,

$$Y^TX\beta = (X^T\beta^TY)^T$$

In this particular equation, it can also be inferred that $Y^TX\beta$ and $X^T\beta^TY$ are both scalars. Therefore, the equation for s can be further simplified as:

$$s = Y^TY - 2\beta^TX^TY + X^T\beta^TX\beta$$

Upon calculating the derivative of this equation and setting it to 0:

$$\frac{\partial s}{\partial \beta^*} = 0 - 2X^TY + 2X^TX\beta^*$$

$$-2X^TY + 2X^TX\beta^* = 0$$

$$2X^TY = 2X^TX\beta^*$$

Therefore, now we can finally prove:

$$X^TY = X^TX\beta^*$$

Sub-Problem 4

We are required to show that

$$\beta * = (X^T X)^{-1}(X^T Y)$$

From sub-problem 2, we computed the sum of squared residuals as being:

$$s = e^T \cdot e = (y_1 - \beta_0 - x_{11}\beta_1 - \cdots - x_{1k}\beta_k)^2 + (y_2 - \beta_0 - x_{21}\beta_1 - \cdots + x_{2k}\beta_k)^2 - \cdots + (y_n - \beta_0 - x_{n1}\beta_1 - \cdots - x_{nk}\beta_k)^2$$

This can be written in a more generalized form as:

$$s = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} X_{ij}\beta_j \right)^2$$

where y_i is the observed value, $\hat{y_i}$ is the predicted value, X_{ij} are the elements of the matrix X, and β_j are the coefficients to estimate.

Upon computing the derivative of this scalar and setting it to zero (so that we can find the minimum), we get:

$$\frac{\partial s}{\partial \beta_j} = -2 \sum_{i=1}^{n} X_{ij}(y_i - \sum_{l=0}^{k} X_{il}\beta_l) = 0, \quad \text{for each } j = 0, 1, ..., k$$

This can be further simplified as:

$$X^T X \beta = X^T Y$$

Here, X^T is the transpose of the matrix X, and Y is the outcome value vector.

To find the vector β^* (which minimizes s), we solve the above equation. Assuming that $X^T X$ is invertible, we multiply both sides by $(X^T X)^{-1}$:

$$\beta^* = (X^T X)^{-1} X^T Y$$

Matrix Algebra - based proof

From sub-problem 3, we were able to prove that:

$$X^T Y = X^T X \beta^*$$

To find the vector β^* (which minimizes s), we solve the above equation. Assuming that $X^T X$ is invertible, we multiply both sides by $(X^T X)^{-1}$:

$$\beta^* = (X^T X)^{-1} X^T Y$$

Problem 2

The Python Code for this problem is titled assignment-02.ipynb and attached with this submission. Upon implementing Linear Regressions with the matrix algebra method and the blackbox scikit-learn method, the following are the results from calculating the coefficients:

Coefficient Number	Matrix Algebra	Blackbox Implementation	delta
0	2.93889	2.93889	-1.95399e-14
1	0.0457646	0.0457646	-9.71445e-17
2	0.18853	0.18853	-1.16573e-15
3	-0.00103749	-0.00103749	3.22442e-16

Coefficient 0 refers to the intercept.

It could be seen from the above results that there is no difference (since it's actually infinitesimally small) between the values of the regression coefficients as calculated by the matrix algebra and blackbox linear regression approaches.